False discovery rate controlling procedures for discrete tests

June 4, 2012

Ruth Heller
Department of Statistics and Operations Research, Tel-Aviv university, Tel-Aviv, Israel. E-mail: ruheller@post.tau.ac.il
Hadas Gur
Faculty of Industrial Engineering and Management, Technion – Israel Institute of Technology, Haifa, Israel. E-mail:gur.hadas@gmail.com

Abstract

Benjamini and Hochberg (1995) proposed the false discovery rate (FDR) as an alternative to the family-wise error rate in multiple testing problems, and proposed a procedure to control the FDR. For discrete data this procedure may be highly conservative. We investigate alternative, more powerful, procedures that exploit the discreteness of the tests and have FDR levels closer in magnitude to the desired nominal level. Moreover, we develop a novel step-down procedure that dominates the step-down procedure of Benjamini and Liu (1999) for discrete data. We consider an application to pharmacovigilance spontaneous reporting systems, that serve for early detection of adverse reactions of marketed drugs.

1 Introduction

In many modern applications the data is discrete, and hundreds, or even thousands, of hypotheses are simultaneously tested. In order to control for false-positives, a multiple testing procedure may be applied that either controls the probability of at least one false positive (family-wise error) or the expected proportion of true null hypotheses rejected out of all rejected hypotheses, known as the false discovery rate (FDR, Benjamini and Hochberg 1995). Researchers have been active in developing methodologies for controlling the FDR, see Benjamini (2010) for an overview. However, little has been written about controlling the FDR when the data is discrete.

In many modern applications, it may be more appropriate to apply a multiple testing procedure that controls the false discovery rate (FDR) over a family-wise error controlling procedure. One such application with discrete data comes from pharmacovigilance systems for marketed medicines, that collect and monitor spontaneous reports of suspected adverse events from health-care providers. In order to detect new adverse drug reactions after the drug marketing approval, multiple hypotheses of no association between drugs and adverse events are simultaneously tested periodically in the pharmacovigilene databases. Pharmacovigilance and all drug safety issues are relevant for everyone whose life is touched in any way by medical interventions. In the analysis of pharmacovigilance systems, the aim is not to substitute the expertise of the pharmacovigilance experts but rather to draw attention to unexpected associations by acting as hypothesis generators. The associations thus found are then destined to be further investigated, so it is possible to tolerate few false discoveries as long as they are a small fraction of the discoveries.

Another modern application area is genomics research. High-throughput next generation sequencing (HT-NGS) technologies output a list of sequence reads. These sequences are mapped to their genomics locations. The data for statistical analysis is tag counts. In order to test for enriched regions, the null distribution of counts is used. Discrete data is also encountered in genome wide association studies, where minor allele frequency of diseased and non-diseased individuals are compared simultaneously in hundreds of thousands of single-nucleotide polymorphisms.

The procedure introduced by Benjamini and Hochberg (1995), henceforth referred to as the BH procedure, is the original and still very popular multiple testing procedure for controlling the FDR. If the null distribution of the p-values is uniform and the p-values are independent, then the FDR of the BH procedure at level q is $\frac{m_0}{m}q$, where m and m_0 are the number of hypotheses and the number of true null hypotheses respectively (Benjamini and Hochberg, 1995). However, for discrete test statistics the null distribution of the p-values is stochastically larger than the uniform, and therefore the FDR of the BH procedure may be much smaller than $\frac{m_0}{m}q$. This is so because the expression for the FDR of the BH procedure involves sums with the terms $Pr_{H_i}(P_i \leq \frac{k}{m}q)$ (Benjamini and Yekutieli, 2001), where P_i is the p-value of a true null hypotheses H_i, $k = 1, \dots, m$. If the null distribution of the p-value is uniform, then $Pr_{H_i}(P_i \leq \frac{k}{m}q) = \frac{k}{m}q$. But for discrete data, $Pr_{H_i}(P_i \leq \frac{k}{m}q)$ may be less than $\frac{k}{m}q$ and, the greater the gaps between $Pr_{H_i}(P_i \leq \frac{k}{m}q)$ and $\frac{k}{m}q$, the smaller the true FDR level of the BH procedure. Thus, the BH procedure may be conservative for discrete data, in the sense that its actual FDR level may be smaller than $\frac{m_0}{m}q$. Note that this conservatism does not go away with an increase in the number of hypotheses, nor with modifications of the original BH procedure that can provide higher power by incorporating an estimate of the number of null hypotheses (such as, for e.g., the adaptive procedure in Benjamini et al. (2006)).

Few other approaches that take the discreteness into account for FDR control have been suggested in the literature. Kulinskaya and Lewin (2009) suggested an FDR controlling procedure using randomized p-values (from randomized tests) to account for the discreteness of the null distribution, thus guaranteeing that the p-values are uniformly distributed under the null, and therefore that the FDR is controlled exactly at the desired level when the p-values are independent. Interpretation of results is not straightforward in this case though, due to the randomness of the p-values. Gilbert (2005) proposed a two step FDR controlling procedure for discrete data. First, remove the null hypotheses with test statistics that are unable to reach a certain level of significance. Second, apply the BH procedure to the remaining hypotheses. This approach does not exploit the discreteness of the test statistics that are not removed in the first step. Heyse (2011) suggested a discrete BH procedure, that exploits more fully the discrete null distributions of the test statistics, and demonstrated in simulations that it has power advantage over the procedure of Gilbert (2005). However, the procedure by Gilbert (2005) controls the FDR while the procedure in Heyse (2011) may be anti-conservative. Ahmed et al. (2010) used $midP$-values in conjunction with an FDR controlling procedure for the analysis of pharmacovigilance systems, and provided simulation results that suggest that it is an improvement over using the p-values.

Our first aim in this work is to study the properties of the BH procedure using $midP$-values. We will prove that the actual FDR level of the BH procedure based on $midP$-values is closer to the nominal level than the BH procedure based on p-values. We will also derive an upper bound on the FDR level of the BH procedure based on $midP$-values. A straightforward modification of the procedure in Gilbert (2005) will be to apply this procedure using $midP$-values. We will compare and contrast this resulting new procedure with the procedure suggested by Heyse (2011).

The BH procedure is a step-up procedures. Benjamini and Liu (1999) suggested a step-down procedure for FDR control, called henceforth the BL procedure. Our second aim in this work is to study discrete analogues to the BL procedure. We develop a novel discrete BL procedure and prove that it controls the FDR at the nominal level. We will compare and contrast this novel procedure with proven FDR control, to the new procedure that results from removing first the null hypotheses with test statistics that are unable to reach the nominal level of significance, and then applies the BL procedure on $midP$ values.

The paper is organized as follows. Section 2 introduces the relevant procedures and discusses theoretical properties of these procedures. Section 3 applies the procedure on an example from a pharmacovigilance database. In the example, more suspect drugs can indeed be discovered with the procedures that take discreteness into account. Section 4 evaluates the proposed procedures by simulation, and section 5 concludes with final remarks. An R package *discreteMTP* to perform the step-up and the step-down discrete multiple testing variates of BH and BL, respectively, is available from the first author web page or CRAN.

Table 1: Table relating treatment to adverse event, for 10 studies. Among the treated, the occurrences and nonoccurrences were X_{11} and X_{12} respectively; among the controls, the occurrences and nonoccurrences were X_{21} and X_{22} respectively; the p-value was computed from a one-sided Fisher's exact test for 2×2 tables; the $midP$-value is the average of the p-value with the next smallest p-value that could possibly be observed in Fisher's exact test with the same fixed margins as observed.

	X_{11}	X_{12}	X_{21}	X_{22}	p-value	$midP$-value
1	1.000	15.000	13.000	3.000	0.000	0.000
2	2.000	36.000	12.000	20.000	0.001	0.000
3	1.000	14.000	7.000	6.000	0.009	0.005
4	10.000	30.000	12.000	8.000	0.009	0.006
5	0.000	20.000	5.000	18.000	0.035	0.017
6	2.000	5.000	7.000	2.000	0.072	0.039
7	8.000	16.000	15.000	12.000	0.095	0.062
8	3.000	11.000	7.000	15.000	0.389	0.267
9	5.000	12.000	5.000	10.000	0.555	0.411
10	7.000	14.000	5.000	20.000	0.914	0.834

2 FDR controlling procedures for discrete data

Consider a family of m hypotheses $H_1, \ldots, H_m$ with corresponding p-values $p_1, \ldots, p_m$. Let I_0 be the indices of the true null hypotheses, and $m_0 = |I_0|$ be the number of null hypotheses. Sorting these p-values, we get $p_{(1)} \leq \ldots \leq p_{(j)} \leq \ldots \leq p_{(m)}$ with corresponding null hypotheses $H_{(1)}, \ldots, H_{(j)}, \ldots, H_{(m)}$.

In this section we illustrate the different FDR controlling procedures using a small data example summarized in Table 1, that relates treatment to an adverse event for 10 studies. This is a subset of the 41 studies considered in Efron (1996).

2.1 The BH procedure on $midP$-values

The $midP$-value was suggested by Lancaster (1961) to replace the p-value in discrete tests. The p-values are made smaller by averaging the actual observed p-value with the next smaller p-value that could possibly be observed. The probability of observing a $midP$-value less than α should better approximate the nominal level α, because the distribution of the $midP$-value under the null hypothesis is closer to uniform than is the P-value. For motivation and theoretical justifications, see Lancaster (1961), Routledge (1994),Berry and Armitage (1995) and Fellows (2010). However, unlike the p-value, a test based on the $midP$-value may exceed the nominal significance level α. Agresti and Gottard (2007) write "..we believe it is more sensible to use a method for which the actual error rate is closer to the nominal error rate than happens with traditional exact inference. Inference based on the $midP$-value is a simple way to achieve

Table 2: The adjusted p-values from the multiple testing procedures on the p-values or $midP$-values of Table 1. The columns from left to right are the adjusted p-values from (1) the BH procedure on the p-values; (2) the BH procedure on the $midP$-values; (3) the DBH procedure; (4) the BL procedure on p-values; (5) the BL procedure on $midP$-values; (6) the DBL procedure.

	BH-	$midP$+BH	DBH	BL	$midP$+BL	DBL
	adjusted	-adjusted	-adjusted	-adjusted	-adjusted	-adjusted
1	0.000	0.000	0.000	0.000	0.000	0.000
2	0.004	0.002	0.001	0.007	0.004	0.002
3	0.023	0.014	0.012	0.054	0.029	0.023
4	0.023	0.014	0.012	0.054	0.029	0.023
5	0.070	0.035	0.038	0.115	0.060	0.077
6	0.119	0.064	0.062	0.155	0.089	0.078
7	0.135	0.089	0.082	0.155	0.091	0.107
8	0.486	0.333	0.351	0.231	0.182	0.200
9	0.617	0.457	0.442	0.231	0.182	0.200
10	0.914	0.834	0.846	0.231	0.182	0.200

this goal." Similarly, when testing simultaneously multiple hypotheses using p-value based multiple testing procedures, we argue that it is more sensible to use $midP$-values in place of p-values.

Since the $midP$-values are smaller than the p-values, using $midP$-values instead of p-values in an FDR controlling procedure will lead to at least as many rejections as the FDR controlling procedure based on p-values. For the small data example summarized in Table 1, the adjusted p-values from a BH procedure on p-values and on $midP$-values are summarized in Table 2. From Table 2 we see that applying the BH procedure at level $q = 0.1$ on p-values and on $midP$-values led to 5 and 7 rejections respectively.

Since the distribution of $midP$ is closer to the uniform than the distribution of the p-value, the true FDR level of FDR controlling procedures will be closer to the nominal level than the true FDR on the original procedures that use p-values. We formalize this statement for the BH procedure on $midP$-values.

Proposition 2.1. *Let $midFDR$ and $origFDR$ be the true FDR levels of the BH procedure using the BH procedure at level q on $midP$-values and on p-values respectively. Then if the p-values are independent,*

$$|\frac{m_0}{m}q - midFDR| \leq \frac{m_0}{m}q - origFDR.$$

Note that the true levels of the BH procedure at level q on $midP$-values and on p-values vary with the probability distributions of the p-values, so a technically more precise statement of the inequality in Proposition 2.1 that does not suppress the dependence of the true FDR levels on the true data generating distributions is

$$|\frac{m_0}{m}q - midFDR(\mathcal{L}(P_1), \ldots, \mathcal{L}(P_m)| \leq \frac{m_0}{m}q - origFDR(\mathcal{L}(P_1), \ldots, \mathcal{L}(P_m),$$

where $\mathcal{L}(P_i)$ denotes the true probability distribution of p-value P_i.

From the above proposition, it follows that applying the BH procedure on $midP$-values will result in an FDR level at most $2\frac{m_0}{m}q - origFDR$. A tighter upper bound on the $midFDR$, that is calculable from the known discrete null distributions, is derived in the following proposition.

Proposition 2.2. *Let* $\epsilon_i = \max_{k \in \{1,\ldots,m\}} \frac{Pr_{H_i}(midP_i \leq kq/m)}{kq/m}$. *If the p-values are independent,* $midFDR \leq \frac{q}{m}\sum_{i \in I_0} \epsilon_i$.

It follows that $midFDR \leq \frac{q}{m}\sum_{i=1}^{m} \epsilon_i$, and this upper bound can be computed from the known null distributions. Note that ϵ_i is bounded above by 2, since for $midP$-values $Pr_{H_i}(midP_i \leq x) \leq 2x - Pr_{H_i}(P_i \leq x)$. Therefore, $midFDR \leq 2q$, but this upper bound may be far from tight. Depending on the exact null distributions, the upper bound $\frac{q}{m}\sum_{i=1}^{m} \epsilon_i$ may be much closer to q than to $2q$.

Incorporating the simple two-step combination of the Tarone and BH procedures in Gilbert (2005) on $midP$-values results in the following procedure.

Procedure 2.1. *The Tarone+midP adjusted BH procedure at level* q *is the following two-step procedure:*

1. *Compute the minimum achievable significance level for hypothesis* H_i*, call it* q_i**. For each* $k = 1, \ldots, m$*, let* $m(k)$ *be the number of hypotheses for which* $q_i* \leq c \cdot q/k$ *(where* $c \geq 1$ *is a predefined constant), and let* K *be the smallest value of* k *such that* $m(k) \leq k$*. Let* $I_K \subset \{1, \ldots, m\}$ *be the set of indices satisfying* $q_i* \leq q/K$*.* I_K *contains the set of* $m(K)$ *indices of hypotheses for which the minimum achievable significance level is below the Bonferroni threshold when testing* $m(K)$ *hypotheses.*

2. *Apply the BH procedure on the midP-values of the family of hypotheses with indices in the set* I_K*.*

The two-step procedure will typically have higher power than the BH procedure on $midP$-values, with the gain in power generally increasing with $m - m(K)$. Propositions 2.1 and 2.2 hold for the Tarone+midP adjusted BH procedure, with m replaced by $m(K)$, since the first step in Procedure 2.1 selects the subset of hypotheses with indices I_K for testing solely based on the null distributions of the p-values, without looking at the realized p-values.

2.2 A discrete step-up procedure

The BH-adjusted p-values Benjamini et al. (2006) are $p_{(j)}^{BHadj} = \min_{i \geq j} \frac{m}{i} p_{(i)}$. The BH procedure at level q is equivalent to rejecting all hypotheses with BH-adjusted p-value $\leq q$. Motivated by this formulation of the BH procedure, Heyse (2011) suggested the following discrete analogue, henceforth called the DBH procedure. The DBH procedure adjusted p-values are

$$p_{(j)}^{DBHadj} = \min_{i \geq j} \frac{\sum_{l=1}^{m} Pr_{H_l}(P_l \leq p_{(i)})}{i}.$$

The DBH procedure at level q is equivalent to rejecting all hypotheses with DBH-adjusted p-value $\leq q$.

The gain from using the DBH procedure over the BH procedure comes from the fact that $Pr_{H_l}(P_l \leq p_{(i)}) \leq p_{(i)}$. If hypothesis H_l cannot achieve a p-value below $p_{(i)}$ then $Pr_{H_l}(P_l \leq p_{(i)}) = 0$ and the dimensionality of the multiple comparisons problem is reduced. If hypothesis H_l can achieve a p-value below $p_{(i)}$ then $Pr_{H_l}(P_l \leq p_{(i)}) \leq p_{(i)}$ and a smaller quantity adds to $p_{(j)}^{DBHadj}$. On the other hand, if all the null distributions are identical then $Pr_{H_l}(P_l \leq p_{(i)}) = p_{(i)}$ and there is no gain in using the DBH procedure over the original BH procedure. Thus follows the proposition below,

Proposition 2.3. *The DBH procedure rejects at least as many null hypotheses as the BH procedure. However, if all null distributions are the same then the DBH procedure rejects exactly the same null hypotheses as the BH procedure.*

For the small example in Table 1, Table 2 shows that adjusted p-values from the DBH procedure are smaller than the adjusted p-values from the BH procedure on p-values, but not necessarily smaller than the adjusted p-values from the BH procedure on $midP$-values.

An example where the DBH procedure does not control the FDR Note that if R hypotheses are rejected by the DBH procedure, then $\frac{\sum_{l=1}^{m} Pr_{H_l}(P_l \leq p_{(R)})}{R} \leq q$, but this does not guarantee that the FDR is controlled, since the FDR is $E(\frac{\sum_{i \in I_0} I[P_i \leq p_{(R)}]}{\max(R,1)})$ and this quantity may be larger than q, as the following example demonstrates. Let P_1 be a p-value with atom at 0.02, 0.045 and 1, and let P_2 be a p-value independent of P_1 with atoms at 0.03, 0.055 and 1. For $m = m_0 = 2$, the FDR is equal to

$$\begin{aligned}
&P(V > 0) = Pr(P_1 = 0.02] + Pr(P_2 = 0.03) \\
&-Pr(P_1 = 0.02) \times Pr(P_2 = 0.03) + Pr(P_1 = 0.045) \times Pr(P_2 = 0.055) \\
&= 0.02 + 0.03 - 0.02 \times 0.03 + 0.025 \times 0.025 = 0.050025.
\end{aligned}$$

2.3 A discrete step-down procedure

The BL procedure (Benjamini and Liu, 1999) is a step-down multiple comparisons procedure for FDR control, so it compares the smallest p-value with the first critical value, and proceeds to compare the second smallest p-value with the second critical value only if the smallest p-value was below its critical value; as soon as a p-value is above its critical value, no further comparisons are made. The critical values in the BL procedure are $\delta_i = 1 - [1 -$

$\min(1, \frac{m}{m-i+1}q)]^{\frac{1}{m-i+1}}, i = 1, \ldots, m$. The procedure find the smallest p-value among all those satisfying $p_{(j)} \leq \delta_j$, call it $p_{(R+1)}$, and reject the R null hypotheses whose p-value is at most $p_{(R)}$. Benjamini and Liu (1999) proved that this procedure controls the FDR at the nominal level q for independent test statistics, and Sarkar (2002) demonstrated that the FDR is controlled also if the test statistics are positive dependent in some sense. Benjamini and Liu (1999) show that the BL procedure neither dominates nor is dominated by the BH procedure.

We consider the following new discrete analogue to the BL procedure, henceforth called the DBL procedure. The DBL procedure will use the following critical values

$$\overset{\sim}{\delta}_i = \max\{\max\{z : \frac{m-i+1}{m}(1 - \prod_{j=i}^{m}(1 - Pr_{H_{(j)}}(P_j \leq z))) \leq q\}, \overset{\sim}{\delta}_{i-1}\}, \quad \overset{\sim}{\delta}_0 = 0.$$

The correspondence between the BL and DBL procedures can best be seen by expressing their respective adjusted p-values. The BL adjusted p-values are

$$p_{(i)}^{BLadj} = \max\{\frac{m-i+1}{m}(1 - (1 - p_{(i)})^{m-i+1}), p_{(i-1)}^{BLadj}\}, \quad p_{(0)}^{BLadj} \doteq 0.$$

The discrete BL adjusted p-values are

$$p_{(i)}^{DBLadj} = \max\{\frac{m-i+1}{m}(1 - \prod_{j=i}^{m}(1 - Pr_{H_{(j)}}(P_j \leq p_{(i)}))), p_{(i-1)}^{DBLadj}\}, \quad p_{(0)}^{DBLadj} \doteq 0.$$

Since $Pr_{H_{(j)}}(P_j \leq p_{(i)}) \leq p_{(i)}$, it follows that $p_{(i)}^{DBLadj} \leq p_{(i)}^{BLadj}$.

Proposition 2.4. *For independent test statistics, the DBL procedure controls the FDR at the nominal level.*

See Appendix C for the proof.

The proposition implies that for independent test statistics, the DBL procedure should always be preferred over the BL procedure with discrete data since it will be uniformly more powerful than the BL procedure and has guaranteed FDR control. For the small example in Table 1, Table 2 shows that adjusted p-values from the DBL procedure are smaller than the adjusted p-values from the BL procedure on p-values, but not necessarily smaller than the adjusted p-values from the BL procedure on *midP*-values.

Relaxation of the independence assumption in Proposition 2.4 For FDR control of the DBL procedure, it is enough to assume that the joint distribution of statistics from true nulls is independent of the joint distribution of statistics from false nulls and that the Sidak inequality is satisfied on the test statistics from true nulls. There is no restriction on the joint dependency of statistics from false nulls. Sidak's inequality (Ge et al., 2003) is $Pr(P_1 \geq p_1, \ldots, P_m \geq p_m) \geq \Pi_{i=1}^{m} Pr(P_i \geq p_i)$.

3 An Example

The Medicines and Healthcare products Regulatory Agency (MHRA, `http://www.mhra.gov.uk/`) in the United Kingdom operate post-marketing surveillance for reporting, investigating and monitoring of adverse drug reactions to medicines and incidents with medical devices. Their database contains complete listings of all suspected adverse drug reactions or side effects, which have been reported by healthcare professionals and patients to the MHRA via the Yellow Card Scheme (`http://yellowcard.mhra.gov.uk/`). The Yellow Card Scheme receives more than 20,000 reports of possible side effects each year. Half a million reports were received in the scheme's first 40 years. In 2007, more than 500 defects related to medicines were reported to the MHRA, resulting in the issue of more than 30 Drug Alert. All reports made to the MHRA on suspected reactions to drugs are listed in the Drug Analysis Prints. We use data from the Drug Analysis Prints for illustration.

To investigate the association between reports of amnesia and suspected drugs, we extracted the number of reported cases of amnesia as well as the total number of adverse events reports for each of the 2466 drug in the database. From the total of 686911 adverse events reports, 2051 contained cases of amnesia. For each drug, the association between the drug and amnesia was tested by a one-sided Fisher's exact test. Specifically, for drug i the 2×2 contingency table for testing for association with amnesia was

	Amnesia	Not Amnesia
Drug i	$A_{11}(i)$	$A_{12}(i)$
Other drugs	$2051 - A_{11}(i)$	$686911 - 2051 - A_{12}(i)$

where $A_{11}(i)$ are the number of Amnesia cases reported for drug i, and $A_{11}(i) + A_{12}(i)$ are the number of cases reported to have adverse events for drug i. Table 3 shows the adjusted p-values from the following 6 procedures: the BH procedure on the p-values and on the $midP$-values respectively, the DBH procedure , the BL procedure on the p-values and on the $midP$-values respectively, and the DBL procedure. The adjusted p-values using the discrete variants of the BH or BL procedure were indeed smaller than the original procedures, and provide more discoveries at a predefined FDR level. Specifically, at the nominal level of $q = 0.05$, the number of drugs discovered to be associated with amnesia by the original BH procedure on p-values was 23, and there were two additional discoveries using the BH procedure on $midP$-values. Applying the DBH procedure provided a total of 27 discoveries.

Should we indeed pay attention to the additional discoveries provided by the discrete step-up procedures over the BH procedure? For one such discovery, Bupropion, the answer is clearly positive. The adjusted p-values for the drug Bupropion by the BH procedure on p-values, on $midP$-values, and by the DBH procedure were, respectively, 0.053, 0.0392, and 0.0134. Therefore, at level $q = 0.05$ this association would not be discovered by the original BH procedure but would be discovered by the two discrete analogues. Evidence that Bupropion is associated with memory disorders in the French pharmacovigilance database Bupropion was reported by Chavant et al. (2011).

Table 3: The 27 smallest adjusted p-values from the multiple testing procedures on the p-values or $midP$-values of the adverse event data in Section 3. The columns from left to right are the adjusted p-values from (1) the BH procedure on the p-values; (2) the BH procedure on the $midP$-values; (3) the DBH procedure; (4) the BL procedure on p-values; (5) the BL procedure on $midP$-values; (6) the DBL procedure.

	BH	BH midPV	DBH	BL	BL mid PV	DBL
BUPROPION	0.0503	0.0392	0.0134	0.6909	0.6003	0.2681
GABAPENTIN	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
INDOMETHACIN	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
LACOSAMIDE	0.0331	0.0176	0.0085	0.5254	0.3270	0.1734
LEVETIRACETAM	0.0054	0.0033	0.0014	0.0958	0.0592	0.0258
LITHIUM	0.0001	0.0001	0.0000	0.0020	0.0012	0.0004
LORAZEPAM	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
MEFLOQUINE	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
MIDAZOLAM	0.0023	0.0013	0.0006	0.0353	0.0202	0.0095
OXCARBAZEPINE	0.1377	0.0752	0.0349	0.9645	0.8569	0.5942
PREGABALIN	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
RIMONABANT	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
SERTRALINE	0.0695	0.0490	0.0186	0.8131	0.6961	0.3621
SIMVASTATIN	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
STRONTIUM RANELATE	0.0066	0.0038	0.0019	0.1222	0.0728	0.0360
TEMAZEPAM	0.0001	0.0001	0.0000	0.0012	0.0007	0.0003
TOPIRAMATE	0.0046	0.0028	0.0012	0.0739	0.0465	0.0206
TRIAZOLAM	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
VARENICLINE	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
VIGABATRIN	0.0050	0.0030	0.0014	0.0854	0.0526	0.0236
ZOLPIDEM	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
ZOPICLONE	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
CITALOPRAM	0.0010	0.0007	0.0002	0.0141	0.0096	0.0032
DEXAMPHETAMINE	0.0073	0.0038	0.0020	0.1398	0.0753	0.0406
ETHANOL	0.1274	0.0695	0.0325	0.9527	0.8243	0.5562
FLUOXETINE	0.0236	0.0176	0.0062	0.3995	0.3251	0.1250
PAROXETINE	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Note that due to the limitations of the way data are gathered into pharmacovigilance systems, neither our analysis, nor the analysis in Chavant et al. (2011), can confirm that Bupropion affects memory, but the analysis does suggest that further investigation into the effect of this drug should be pursued.

In fact, for all the drugs discovered by our analysis, further investigations are necessary for establishing whether the suspect drugs indeed cause amnesia. These further investigations may be costly, and therefore it is important not to launch investigations into too many false leads. The discrete procedures have higher power to discover true associations between drug and amnesia than the original procedures, while being careful to guarantee that only a small fraction of the discovered associations may be false positives, thus these procedures suit well the purpose of pharmacovigilance systems.

4 A simulation study

The power and FDR level of the various procedures are compared in simulations. We examined simulation settings similar to the settings in Gilbert (2005), except that we examine one-sided tests rather than the two-sided tests considered in Gilbert (2005).

4.1 Structure of the simulation

A vector of $m = 100$ binary responses is observed for each of N individuals in each group, and the goal is to test simultaneously the m hypotheses $H_i : p_{1i} = p_{2i}, i = 1, \ldots, m$, where p_{ij} is the success probability for the ith binary response in group j ($i \in \{1, \ldots, m\}$ and $j \in \{1, 2\}$). For fraction f_1 and f_2 of the m hypotheses, the null was true with success probability 0.01 and 0.10 respectively. The remaining null hypotheses were false with success probabilities 0.10 and 0.30. As f_1 increases, procedures that consider first a Tarone adjustment, as detailed in step 1 of Procedure 2.1, and then an FDR controlling procedure, have more power than just the FDR controlling procedure. For each data set an unadjusted p-value from Fisher's exact test is computed for each of the m positions at which there is at least one success in the pooled data set. For independent test statistics, the data for each of the m contingency tables was independently generated. For dependent test statistics the data was generated as follows. For each of the N subjects in group s, $s \in \{1, 2\}$, a multivariate normal outcome vector $\mathbf{X}_{si} \sim N_m(\mu_s, \mathbf{\Sigma})$ was first generated, and $\mathbf{Y}_{si} = I[\mathbf{X}_{si} < 0]$. The parameter vector $\mu_{\mathbf{s}}$ was chosen to reflect probabilities of 0.1 or 0.3. The off-diagonals in the covariance $\mathbf{\Sigma}$ received the value $\rho \in \{0, 0.01, 0.5, 0.9\}$.

The 4 new procedures considered are the DBH procedure, the Tarone+*midP* adjusted BH procedure, the DBL procedure, and the Tarone+*midP* adjusted BL procedure. They were compared to the BH and BL procedures.

The FDR level of each procedure is the fraction of rejected hypotheses that are truly null out of all hypotheses rejected, averaged over the 1000 simulations.

Table 4: The average realized FDR (SE) over 1000 simulations for various sample sizes N in each group, of a simulation study with $m = 20$ hypothesis, out of which 4 hypotheses are known with success probability 0.01, 15 hypotheses are null with success probability 0.1, and 1 hypothesis is non-null with success probabilities 0.1 and 0.3. The rows are the results for (1) the DBH procedure; (2) the BH procedure on the $midP$-values; (3) the DBH procedure on the p-values; (4) the DBL ; (5) the BL procedure on $midP$-values; (6) the BL procedure on the p-values.

Procedure	N=25	N=50	N=75	N=100	N=125	N=150	N=175	N=200
DBH	0.036 (0.005)	0.029 (0.004)	0.056 (0.006)	0.054 (0.005)	0.049 (0.005)	0.051 (0.005)	0.054 (0.005)	0.037 (0.004)
Tarone+$midP$ BH	0.008 (0.003)	0.024 (0.004)	0.035 (0.005)	0.046 (0.005)	0.047 (0.005)	0.043 (0.005)	0.051 (0.005)	0.036 (0.004)
BH	0.001 (0.001)	0.009 (0.003)	0.015 (0.003)	0.024 (0.004)	0.025 (0.004)	0.019 (0.003)	0.031 (0.004)	0.022 (0.003)
DBL	0.030 (0.005)	0.022(0.004)	0.035 (0.005)	0.038 (0.005)	0.033 (0.004)	0.026 (0.004)	0.034 (0.004)	0.022 (0.003)
Tarone+$midP$ BL	0.009 (0.003)	0.019 (0.004)	0.021 (0.004)	0.028 (0.004)	0.031 (0.004)	0.024 (0.003)	0.032 (0.004)	0.022 (0.003)
BL	0.001 (0.001)	0.007 (0.003)	0.010 (0.003)	0.018 (0.004)	0.017 (0.003)	0.011 (0.002)	0.016 (0.003)	0.012 (0.003)

The power of each procedure is the fraction of non-null hypotheses that are rejected out of all non-null hypotheses, averaged over the 1000 simulations.

4.2 Results of the simulation

Tables 4 and 5 show the resulting FDR and power of the 6 procedures, when applied to independent test statistics from 20 hypotheses. The 20 hypotheses included: one false hypothesis, with success probabilities 0.1 in one group and 0.3 in the other group; four true null hypotheses, with success probability 0.01 in both groups; 15 true null hypotheses, with success probability 0.1 in both groups. Examination of the two tables leads to the following two conclusions, that were true in all simulation settings we considered. First, the discrete procedures had higher FDR levels, but still below the nominal 0.05 level, and were more powerful, than their non-discrete analogues (BH or BL). Second, the power advantages were larger for smaller sample sizes N, since the data was more discrete for smaller N.

Moreover, in table 5 the discrete step-up and discrete step-down procedures are comparable in terms of power, and the FDR level of the DBL procedure is lower than the FDR level of the DBH procedure. For example, for $N = 75$ the FDR of the DBH procedure is estimated to be 0.056 ± 0.006, whereas the FDR of the DBL procedure is estimated to be 0.035 ± 0.005. However, the average power of both procedures is estimated to be 0.71 ± 0.01. The average power of the BH and BL procedures is notably lower, 0.55 ± 0.01. The midP+Tarone adjusted procedures have estimated average power of 0.65 ± 0.02. However, as the number of hypotheses increases (and most of the hypotheses are null), the discrete step-up procedures tend to outperform the discrete step-down procedure, as we show next for $m = 100$.

Figure 1 display the results for independent test statistics in a configuration where 20 and 75 of the hypotheses are null with success probability 0.01 and 0.10 respectively, and 5 of the hypotheses are non-null with success probabilities

Table 5: The average power (SE) over 1000 simulations for various sample sizes N in each group, of a simulation study with $m = 20$ hypothesis, out of which 4 hypotheses are known with success probability 0.01, 15 hypotheses are null with success probability 0.1, and 1 hypothesis is non-null with success probabilities 0.1 and 0.3. The rows are the results for (1) the DBH procedure; (2) the BH procedure on the $midP$-values; (3) the DBH procedure on the p-values; (4) the DBL ; (5) the BL procedure on $midP$-values; (6) the BL procedure on the p-values.

Procedure	N=25	N=50	N=75	N=100	N=125	N=150	N=175	N=200
DBH	0.246 (0.014)	0.454 (0.016)	0.711 (0.014)	0.842 (0.012)	0.918 (0.009)	0.970 (0.005)	0.983 (0.004)	0.991 (0.003)
Tarone+$midP$ BH	0.209 (0.013)	0.412 (0.016)	0.654 (0.015)	0.817 (0.012)	0.912 (0.009)	0.963 (0.006)	0.982 (0.004)	0.988 (0.003)
BH	0.088 (0.009)	0.311 (0.015)	0.552 (0.016)	0.757 (0.014)	0.849 (0.011)	0.945 (0.007)	0.970 (0.005)	0.985 (0.004)
DBL	0.236 (0.013)	0.454 (0.016)	0.709 (0.014)	0.840 (0.012)	0.919 (0.009)	0.970 (0.005)	0.983 (0.004)	0.991 (0.003)
Tarone+$midP$ BL	0.211 (0.013)	0.410 (0.016)	0.652 (0.015)	0.812 (0.012)	0.908 (0.009)	0.965 (0.006)	0.982 (0.004)	0.988 (0.003)
BL	0.088 (0.009)	0.329 (0.015)	0.552 (0.016)	0.751 (0.014)	0.850 (0.011)	0.946 (0.007)	0.972 (0.005)	0.985 (0.004)

0.1 and 0.3. The FDR level (top row) is below the nominal 0.05 level for all procedures, but much closer to 0.05 for the discrete procedures over their non-discrete analogues. The average power is displayed in the bottom row. The first and second columns consider respectively, the step-up and step-down procedures. From examination of the first column, we see that the BH procedure has the lowest FDR level and the lowest average power. The DBH has the highest power and the midP+Tarone procedure is a close second. As the sample size increases, the gain in power from using the discrete procedures is diminished, and at $N = 200$ all procedures have the same power. However, for fixed N the gap in power between the BH procedure and the discrete procedures is similar for $m = 20$ and for $m = 100$. From examination of the second column, we see that the correspondence between the discrete procedures and the BL procedures are very similar to those found for the BH procedure. The FDR level of the BL procedure is very low even for moderate sample sizes, and at $N = 200$ there is still a gap between the discrete procedures and the BL procedure. Finally, looking at the power across the two columns, we see that the step-up procedures are more powerful than the step-down procedures.

Table 6 shows the average power of the procedures considered for a different fraction of true null hypotheses: either 80% or 95% of the hypotheses are null. For all procedures, the power was larger when the fraction of true null hypothesis was smaller. The power advantage of the procedures that adjust for discreteness over the BH or BL procedures was larger in the more difficult situations were the fraction of true null hypotheses was smaller. For example, the DBH procedure was $0.80/0.71 = 1.13$ times larger than the BH procedure for $m_0 = 0.95$, but only $0.91/0.86 = 1.06$ times larger than the BH for $m_0 = 0.80$.

Incorporating dependency among the test statistics, as described in Section 4.1, did not result in any of the procedures being anti conservative. As in the independence setting, the discrete procedures were more powerful (not shown). The BH procedure on p-values appears to control the FDR in more circumstances that are not highly artificial Yekutieli (2008), (Romano et al., 2008).

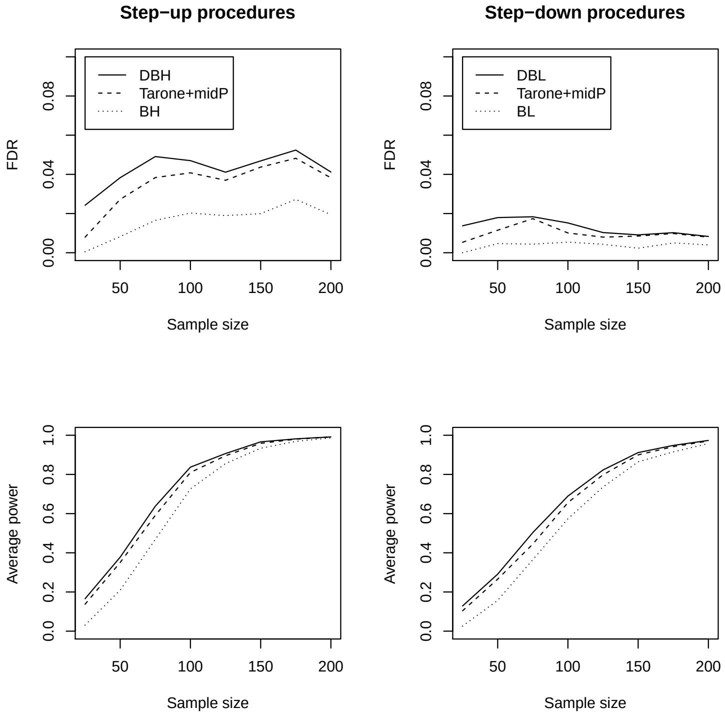

Figure 1: The FDR level (top row) and average power (bottom row) vs. sample size, for the 3 step-up procedures (first column) and the 3 step-down procedures (second column). The number of hypotheses is $m = 100$. 5 hypotheses are non-null with success probabilities 0.1 and 0.3. The success probabilities of the null hypotheses were 0.01 for 20 hypotheses, and 0.10 for 75 hypotheses.

Table 6: The average power (SE) over 1000 simulations of a simulation study with $m = 100$ hypotheses, $N = 100$ subjects in each group, for which the null hypotheses have success probabilities 0.1, and the non-null hypotheses have success probabilities 0.1 and 0.3. The columns vary in the number of true null hypotheses: 80 null hypotheses in the first column, and 95 null hypotheses in the second column. The rows are the results for (1) the DBH procedure; (2) the BH procedure on the $midP$-values; (3) the DBH procedure on the p-values; (4) the DBL ; (5) the BL procedure on $midP$-values; (6) the BL procedure on the p-values.

Procedure	$m_0 = 80$	$m_0 = 95$
DBH	0.909 (0.002)	0.800 (0.006)
Tarone+$midP$ adjusted BH	0.895 (0.002)	0.772 (0.007)
BH	0.861 (0.003)	0.713 (0.007)
DBL	0.675 (0.003)	0.674 (0.007)
Tarone+$midP$ adjusted BL	0.644 (0.003)	0.616 (0.007)
BL	0.586 (0.004)	0.559 (0.007)

This robustness property appears in our simulations to be carried over to the discrete analogues of the BH procedure. Similarly, the step-down procedures were robust to deviations from independence, as considered in our simulations.

5 Summary

We demonstrated that the FDR level may be much lower than the nominal level q when applying the BH or BL procedures at level q on discrete test statistics. By adjusting for discreteness, it was possible to achieve tighter control of the FDR and higher power.

In the simulations considered, the DBH and DBL procedures were more powerful than the Tarone+$midP$ adjusted BH and BL procedures respectively. However, there are important situations where the Tarone+$midP$ adjusted procedures are more powerful. Specifically, when all the null hypotheses are identical, the DBH and DBL procedures are identical to the BH and BL procedures respectively, yet the Tarone+$midP$ adjustment are more powerful. Just as Westfall and Wolfinger (1997) have it for the discrete Bonferroni, the gain in power of the DBH and DBL procedures comes from the fact that the tests are not identically distributed under the null.

Acknowledgements

We thank Felix Abramovich, Yoav Benjamini, Abba Krieger and Daniel Yekutieli for useful discussions. Moreover, we thank Abba Krieger for providing the example where the DBH procedure does not control the FDR in Section 2.2.

References

Agresti, A. and Gottard, A. (2007). Nonconservative exact small-sample inference for discrete data. *Computational statistics & Data analysis*, 51:6447–6458.

Ahmed, I., Dalmasso, C., Harmaburu, F., Thiessard, F., Broet, P., and Tubert-Bitter, P. (2010). False discovery rate estimation for frequentist pharmacovigilance signal detection methods. *Biometrics*, 66:301–309.

Benjamini, Y. (2010). Discovering the false discovery rate. *Journal of the Royal Statistical Society*, 72 (4):405–416.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate - a practical and powerful approach to multiple testing. *J. Roy. Stat. Soc. B Met.*, 57 (1):289–300.

Benjamini, Y., Krieger, A. M., and Yekutieli, D. (2006). Adaptive linear step-up false discovery rate controlling procedures. *Biometrika*, 93 (3):491–507.

Benjamini, Y. and Liu, W. (1999). A step-down multiple hypotheses testing procedure that controls the false discovery rate under independence. *Statistical planning and inference*, 82:163–170.

Benjamini, Y. and Yekutieli, Y. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29 (4):1165–1188.

Berry, G. and Armitage, P. (1995). Mid-p confidence intervals: A brief review. *Journal of the Royal Statistical Society. Series D (The Statistician).*, 44 (4):417–423.

Chavant, F., Favreliere, S., Chebassier, C., Plazanet, C., and Perault-Pochat, M. (2011). Memory disorders associated with consumption of drugs: updating through a case/noncase study in the french pharmacovigilance database. *British Journal of Clinical Pharmacology*, 72 (6):898–904.

Efron, B. (1996). Empirical bayes methods for combining likelihoods. *Journal of the American Statistical Association*, 91 (434):538–550.

Fellows, I. (2010). The minimaxity of the mid p-value under linear and squared loss functions. *Communications in Statistics - Theorey and Methods*, 40 (2):244–254.

Ge, Y., Dudoit, S., and Speed, T. (2003). Resampling-based multiple testing for microarray data analysis. *Test*, 12(1):1–77.

Gilbert, P. (2005). A modified false discovery rate multiple-comparisons procedure for discrete data, applied to human immunodeficiency virus genetics. *Royal Statistical Society , series C*, 54(1):143–158.

Heyse, J. (2011). A false discovery rate procedure for categorical data. In *Recent Advances in Biostatistics: False Discovery Rates, Survival Analysis, and Related Topics*, pages 43–58.

Kulinskaya, E. and Lewin, A. (2009). On fuzzy familywise error rate and false discovery rate procedures for discrete distributions. *Biometrika*, 96(1):201–211.

Lancaster, H. (1961). Significance tests in discrete distributions. *Journal of the American Statistical Association*, 56 (294):223–234.

Romano, J., Shaikh, A., and Wolf, M. (2008). Rejoinder on: Control of the false discovery rate under dependence using the bootstrap and subsampling. *Test*, 17 (3):461–471.

Routledge, R. (1994). Practicing safe statistics with the mid-p. *The Canadian Journal of Statistics*, 22 (1):103–110.

Sarkar, S. (2002). Some results on false discover rate in stepwise multiple testing procedures. *The Annals of Statistics*, 30 (1):239 – 257.

Westfall, P. and Wolfinger, R. (1997). Multiple tests with discrete distributions. *The American Statistician*, 51(1):3–8.

Yekutieli, D. (2008). Comments on: Control of the false discovery rate under dependence using the bootstrap and subsampling. *Test*, 17 (3):458–460.

A Proof of Proposition 2.1

Proof. Let $a_x(i)$ and $b_x(i)$ be the atoms (were an atom is a value the discrete p-value can receive) of the ith discrete p-value just below and just above $x \in (0,1)$ for null distribution i, and let $I[\cdot]$ be the indicator function. Then

$$\begin{aligned}&Pr(midP_i \leq kq/m) = \\ &a_{kq/m}(i) + [b_{kq/m}(i) - a_{kq/m}(i)]I[\frac{b_{kq/m}(i) + a_{kq/m}(i)}{2} \leq kq/m] \qquad (1)\end{aligned}$$

Let $C_k^{(i)}$ denote the event that exactly $k-1$ hypotheses are rejected by the BH procedure on $midP$-values along with true null hypothesis H_i. The FDR level of this procedure may be expressed as follows

$$\begin{aligned}
&midFDR = \sum_{i\in I_0}\sum_{k=1}^{m}\frac{1}{k}Pr(midP_i \le kq/m)Pr(C_k^{(i)}) \\
&= \sum_{i\in I_0}\sum_{k=1}^{m}\frac{1}{k}(a_{kq/m}(i) + [b_{kq/m}(i) - a_{kq/m}(i)]I[\frac{b_{kq/m}(i) + a_{kq/m}(i)}{2} \le kq/m])Pr(C_k^{(i)}) \\
&= origFDR + \sum_{i\in I_0}\sum_{k=1}^{m}\frac{1}{k}[b_{kq/m}(i) - a_{kq/m}(i)]I[\frac{b_{kq/m}(i) + a_{kq/m}(i)}{2} \le kq/m]Pr(C_k^{(i)})
\end{aligned} \tag{2}$$

where $origFDR = \sum_{i\in I_0}\sum_{k=1}^{m}\frac{1}{k}a_{kq/m}(i)Pr(C_k^{(i)})$ is the FDR level of the BH procedure on the p-values. Since $b_{kq/m} - a_{kq/m} \ge 0$, it follows that $midFDR \ge origFDR$. it remains to show that $midFDR \le 2\frac{m_0}{m}q - origFDR$. We will use the following simple lemma:

Lemma A.1. *If* $(b_x(i) + a_x(i))/2 \le x$*, then* $b_x(i) - a_x(i) \le 2x - 2a_x(i)$

Proof. The result is immediate form $(b_x(i) + a_x(i))/2 \le x \iff b_x(i) \le 2x - a_x(i)$. □

Applying the lemma, it follows from equation (2) that $midFDR \le origFDR + 2\frac{m_0}{m}q - 2origFDR$, so the result follows. □

B Proof of Propostion 2.2

Since $\epsilon_i = \max_k \frac{a_{kq/m}(i) + [b_{kq/m}(i) - a_{kq/m}(i)]I[\frac{b_{kq/m}(i)+a_{kq/m}(i)}{2} \le kq/m]}{kq/m}$, then :

$$\begin{aligned}
&midFDR = \sum_{i\in I_0}\sum_{k=1}^{m}\frac{1}{k}(a_{kq/m}(i) + [b_{kq/m}(i) - a_{kq/m}(i)]I[\frac{b_{kq/m}(i) + a_{kq/m}(i)}{2} \le kq/m])Pr(C_k^{(i)}) \\
&= \sum_{i\in I_0}\sum_{k=1}^{m}\frac{1}{k}\frac{kq}{m}\frac{a_{kq/m}(i) + [b_{kq/m}(i) - a_{kq/m}(i)]I[\frac{b_{kq/m}(i)+a_{kq/m}(i)}{2} \le kq/m]}{kq/m}Pr(C_k^{(i)}) \\
&\le \sum_{i\in I_0}\epsilon_i\sum_{k=1}^{m}\frac{q}{m}Pr(C_k^{(i)}) \\
&= \frac{q}{m}\sum_{i\in I_0}\epsilon_i
\end{aligned} \tag{3}$$

C Proof of Proposition 2.4

Proof. First, if $m_0 = 0$ then FDR=0 since there are no false rejections. Second, if $m_0 = 0$ then the only cut-off of interest is $\overset{\sim}{\delta}_1 = \max\{z : 1 - \prod_{i=1}^{m}(1 - Pr_{H_i}(P_i \le$

$z)) \leq q\}$, therefore

$$FDR = FWER = Pr(P_{(1)} \leq \widetilde{\delta}_1) = 1 - \prod_{i=1}^{m}(1 - Pr_{H_i}(P_i \leq \widetilde{\delta}_1)) \leq q$$

where the last inequality follows from the definition of $\widetilde{\delta}_1$.

It remains to prove the proposition for $0 < m_0 < m$. Let $m_1 = m - m_0 > 0$, and let I_1 and I_0 be the index sets of m_1 and m_0 p-values corresponding to the false null and true null hypotheses respectively. Let $\mathbf{P'} = (P'_1, \ldots, P'_{m_1})$ be the p-values corresponding to false null hypotheses.

We will show that $E(Q|\mathbf{P'}) \leq q$, from which the proposition clearly follows. Let $P'_{(1)} \leq \ldots \leq P'_{(m)}$ be the sorted p-values corresponding to false null hypotheses. Let $S \in \{0, \ldots, m_1\}$ be the largest integer i satisfying $P'_{(1)} \leq \widetilde{\delta}_1$ $, \ldots, P'_{(i)} \leq \widetilde{\delta}_i$, where $S = 0$ if $P'_{(1)} > \widetilde{\delta}_1$. Note that $R \geq S{+}V$ since $\widetilde{\delta}_1 \leq \ldots \leq \widetilde{\delta}_m$.

Now we have

$$E(Q|\mathbf{P'}) = E(\frac{V}{R} I[V > 0]|\mathbf{P'}) \leq E(\frac{V}{V + S} I[V > 0]|\mathbf{P'}) \leq \frac{m_0}{S + m_0} Pr(V > 0|\mathbf{P'}) \tag{4}$$

For an index set K of null hypotheses, define the constant

$$c_{q,k} = \max\{z : 1 - \prod_{i \in K}(1 - Pr_{H_i}(P_i \leq z)) \leq q.$$

One can easily verify that for $K \subset K'$ we have $c_{q,K} > c_{q,K'}$.

Let $r'_1, \ldots, r'_S$ be the indices in the vector of p-values of the S smallest p-values corresponding to false null hypotheses. Then

$$Pr(V > 0|\mathbf{P'}) = Pr(\min_{i \in I_0} P_i \leq c_{q\frac{m}{m-S},\{1,\ldots,m\}/\{r'_1,\ldots,r'_S\}}) \tag{5}$$

Since $\{r'_1, \ldots, r'_S\} \subset I_1$ then $I_0 \subset \{1, \ldots, m\}/\{r'_1, \ldots, r'_S\}$. Therefore, $c_{q\frac{m}{m-S},\{1,\ldots,m\}/\{r'_1,\ldots,r'_S\}} \leq c_{q\frac{m}{m-S},I_0}$. If follows that

$$Pr(V > 0|\mathbf{P'}) \leq Pr(\min_{i \in I_0} P_i \leq c_{q\frac{m}{m-S},I_0}) \leq q\frac{m}{m - S} \tag{6}$$

Combining equations (4) and (6) we have

$$E(Q|\mathbf{P'}) \leq \frac{m_0}{S + m_0}\frac{m}{m - S} q \leq q \tag{7}$$

where the last inequality follows from the fact that $m_0 + S \leq m$.

□